DATA SCIENCE CAPSTONE PROJECT 1:

Supplier Pricing Prediction & Segmentation

Content:

1. Code for the Project:

   https://github.com/psanghal/Springboard-Data-Science/blob/master/Capstone%20Project%201/Project%20Notebook%20%26%20Data%20/Capstone1-Final%20Review%20Version-3%20.ipynb

2. Presentation Slides Deck:

   https://drive.google.com/file/d/1-zFFjTEGag_4drCS62M_TlhRuiu2C87C/view?usp=sharing

3. Consolidated Report: (See Enclosed)

Prepared by:

Prashant Sanghal

1. Proposal with problem statement:

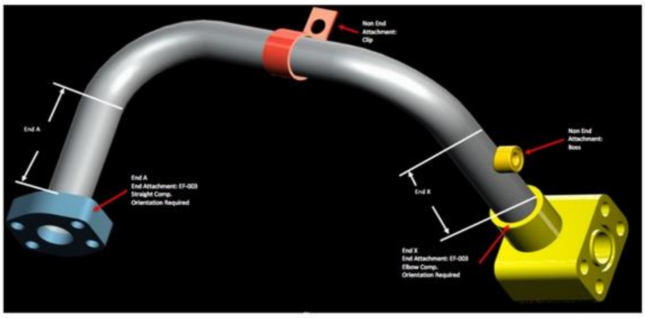
Caterpillar (construction equipment manufacturer) relies on a variety of suppliers to manufacture tube assemblies for their heavy equipment. These assemblies are required in their equipment to lift, load and transport heavy construction loads. We are provided with detailed tube specifications, components, and annual volume datasets. Our goal is to build and train a model that can predict how much a supplier will quote for a given tube assembly based on given supplier pricing, and use this information to further categorize assemblies and suppliers such that any movement in business criteria example recency, frequency, total spend, supplier rebates and more can be accurately classified and responded with by applying appropriate supplier strategy.

To solve this problem, project is divided in two section in the Jupyter Notebook:

- Predicting Supplier Price.
- Categorizing Assembly & Suppliers.

2. Data collection and wrangling summary

This data set was obtained from a public repository, Kaggle. It contained 21 tables detailing various features of the tube assembly. After initial EDA, it was observed that 77% of the values in the dataset were missing values, which after data treatment and feature engineering was reduced to 21,095 observations and 42 attributes. The main focus while data pre-processing was to preserve relevant features in the dataset as well as minimize information loss in the pre-modelled data.



Furthermore, each supplier had their own unique pricing model for tube assemblies which could vary across a number of dimensions, including base materials, number of bends, bend radius, bolt patterns, and end types. Altering any of these specifications, that lower assembly complexity and reduce manufacturing steps while meeting business requirements, would not only result in lower total cost but also technically compliant and reliable component for Caterpillar's construction equipment.

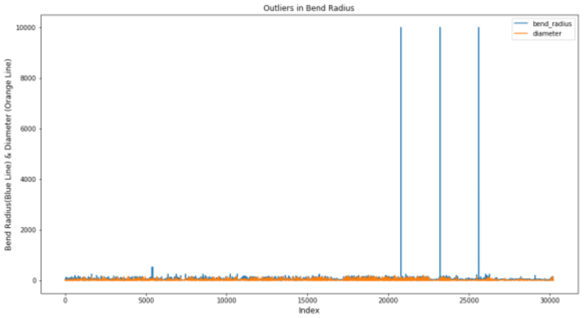
Data Cleaning Steps: This process involved getting data ready for modeling. It required converting dates in to date-time objects and 1000+ categorical strings into binary values (0,1); fixing cost per unit error by including minimum order quantity in supplier pricing; removing duplicates and infinity numbers '9999'; imputing missing values in weight and preserving assembly part name, unique features and more using backfill and forward fill methods. Outliers in tube's bend radius were treated as per design handbook to maintain tube's integrity, while outliers in weight (heavy tube assemblies) were left untreated to maximize information gain.

Here are some visualizations showing before and after treatment results:

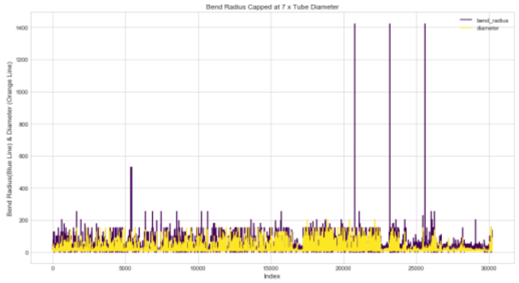
<table>
<tr>
<td>

- Outliers in Bend Radius (Before: 9999)

</td>
<td>Outliers in Bend Radius (After Modification: Capped at 1400)</td>
</tr>
<tr>
<td>



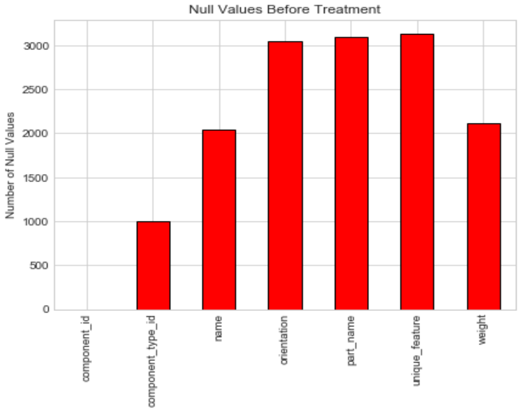
</td>
<td>



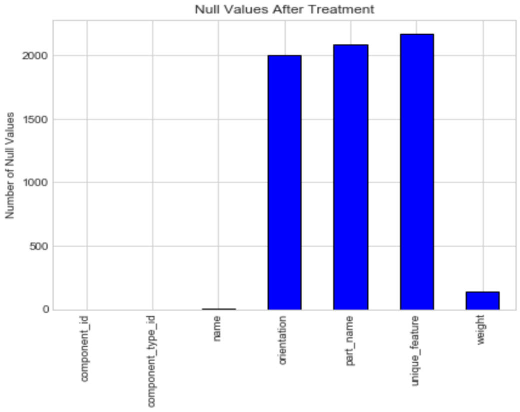
</td>
</tr>
<tr>
<td>

- Null values (Before: Range 0 to 3000)

</td>
<td>Null Values (After Fill Method: Range 0 to 2500)</td>
</tr>
<tr>
<td>

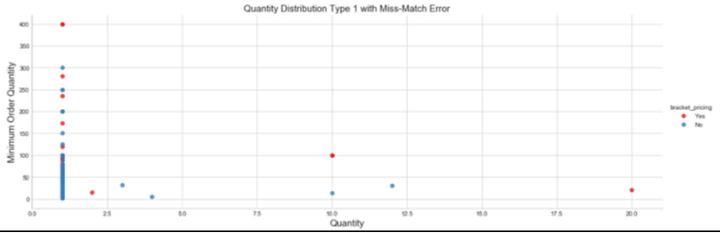


</td>
<td>



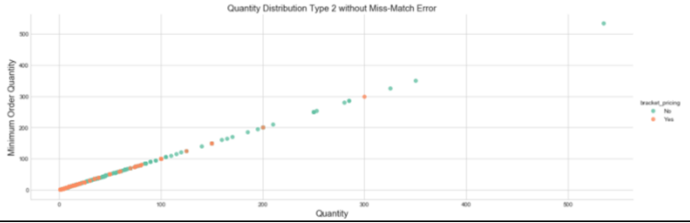
</td>
</tr>
<tr>
<td>

- Cost per quantity miss-match error (Before-Scattered)



</td>
<td>

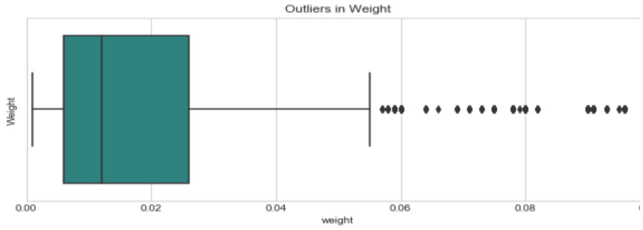
Cost per quantity without miss-match error (After-Linear)



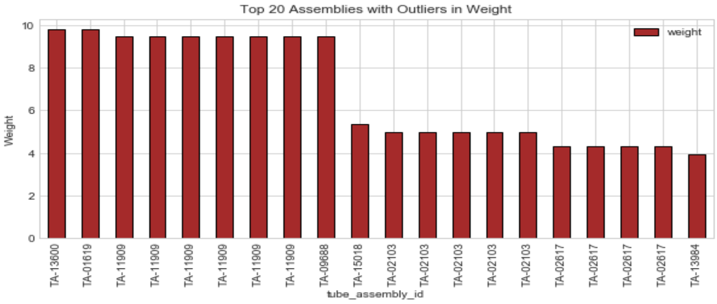
</td>
</tr>
<tr>
<td colspan="2">

- Outliers in Weight (Heavy Assemblies without Treatment)

</td>
</tr>
<tr>
<td>



</td>
<td>



</td>
</tr>
</table>

Finally, I used pandas merge and concatenation function to consolidate various tables into a single table and added new features such as total cost, tube area, supplier rebates for predictive modeling and criteria for category segmentation such as supplier preference, market difficulty, length of contractual relationship and more to benchmark assembly and supplier clusters that met business needs.

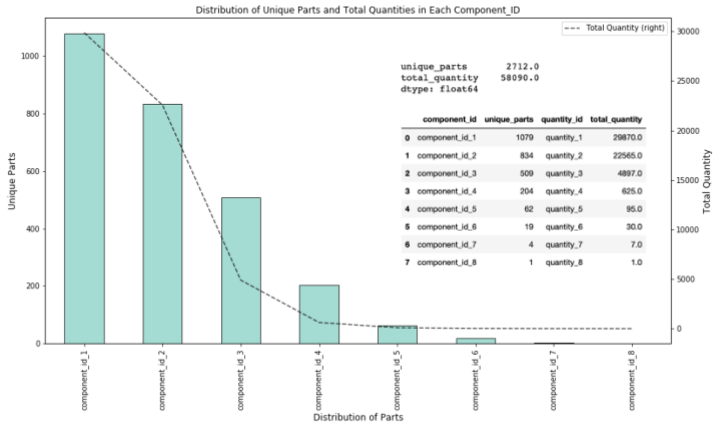
3. Exploratory data analysis summary (visualization and inferential statistics):

Goal of EDA was to uncover data story and understand which components and how many pieces were used historically to manufacture these assemblies.

Bill of materials(BOM):

By exploring (BOM), it was observed that 2,712 unique parts, 8 different components and 58,090 pieces were used in the manufacturing of these assemblies, out of which component_id 1 and component_id 2 had the maximum usage and non-missing values available than the rest of the tube assemblies. Hence, component_id_3 and beyond were rejected to reduce data noise and a new feature 'tube_comp_qty' was added to avoid information loss by identifying tubes that had more than two components.

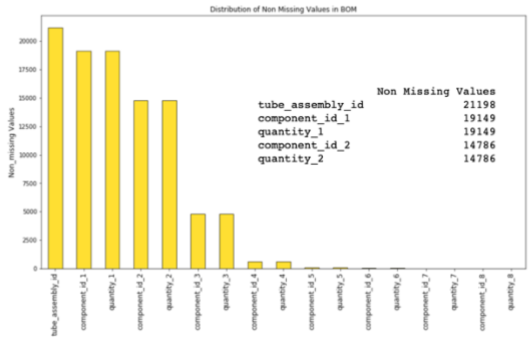
- Breakdown of assembly components



```
unique_parts       2712.0
total_quantity    58090.0
dtype: float64
```

	component_id	unique_parts	quantity_id	total_quantity
0	component_id_1	1079	quantity_1	29870.0
1	component_id_2	834	quantity_2	22565.0
2	component_id_3	509	quantity_3	4897.0
3	component_id_4	204	quantity_4	625.0
4	component_id_5	62	quantity_5	95.0
5	component_id_6	19	quantity_6	30.0
6	component_id_7	4	quantity_7	7.0
7	component_id_8	1	quantity_8	1.0

Distribution of non-missing values in BOM



```
                   Non Missing Values
tube_assembly_id                21198
component_id_1                  19149
quantity_1                      19149
component_id_2                  14786
quantity_2                      14786
```

Distribution of Supplier Assembly Quotes:

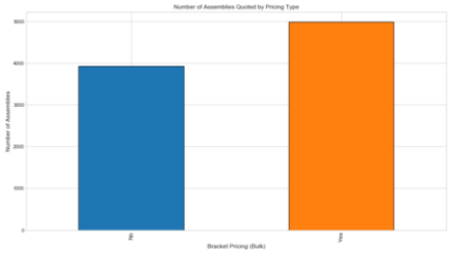
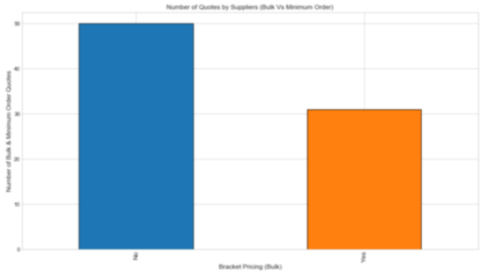
Suppliers provided two types of contract pricing:

I. Bracket or Bulk Pricing which was based on order quantities purchased by Caterpillar, historically.

II. Non-Bracket or Non-Bulk Pricing which was based on minimum order quantities set by the suppliers as a threshold for accepting Caterpillar's purchase order.

From business perspective, suppliers who provided both types of contract pricing were likely to carry higher assortment of various assemblies as well as could respond better to business needs by customizing orders than sole suppliers, who single/fewer assembly types. In the notebook, I compared two pricing types and noticed that 24 out of 57 suppliers offered both contract pricing and had quoted non-bulk assemblies more often than the bulk assemblies but never quoted both pricing for the same assembly i.e. if a supplier quoted bulk price for the assembly, non-bulk pricing was not provided.

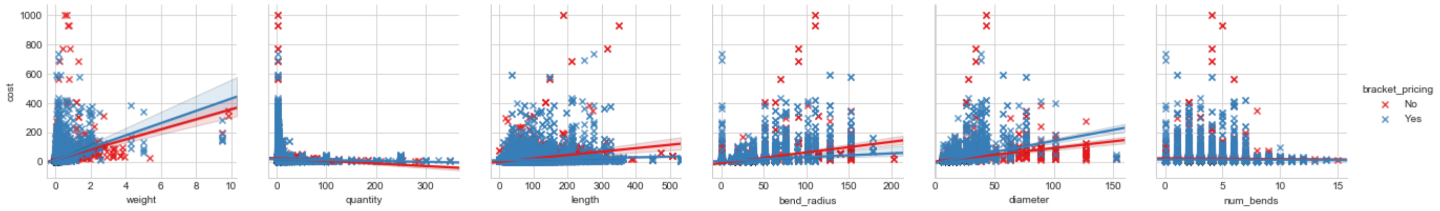
Supplier Pricing Type Distribution:

Sole suppliers quoted more bulk assemblies than non-bulk. 	Multi-  pricing suppliers quoted more non-bulk assemblies than bulk

Supplier Assembly Cost Dependency:

Supplier price was found dependent on various features such as tube specifications, type of contract pricing and order quantities procured in the past. Caterpillar, also provided with annual usage which was a useful feature in anticipating demand for the next year. Here are some visualizations showing relationship between cost and assembly features.

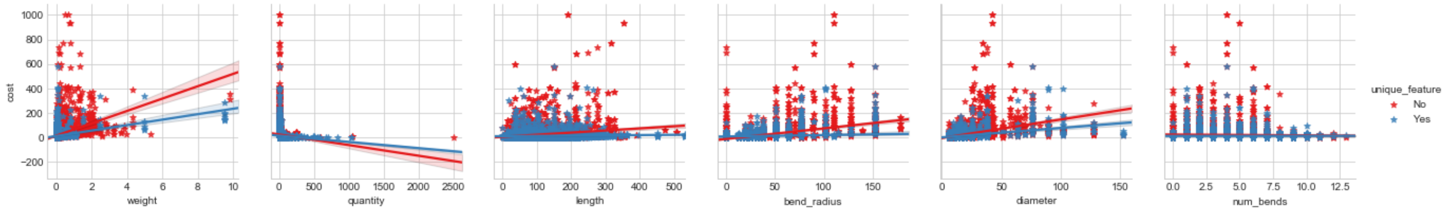
- Cost Vs Tube Specification by Bracket Pricing:



OBSERVATION:

1. Cost increases with increase in weight/length/bend_radius/diameter.
2. Cost decreases with increase in order quantity.
3. Heavier and bigger diameter assemblies when purchased in bulk result in higher cost increase.
4. Cost is not affected by the increase in number of bends in the tube.

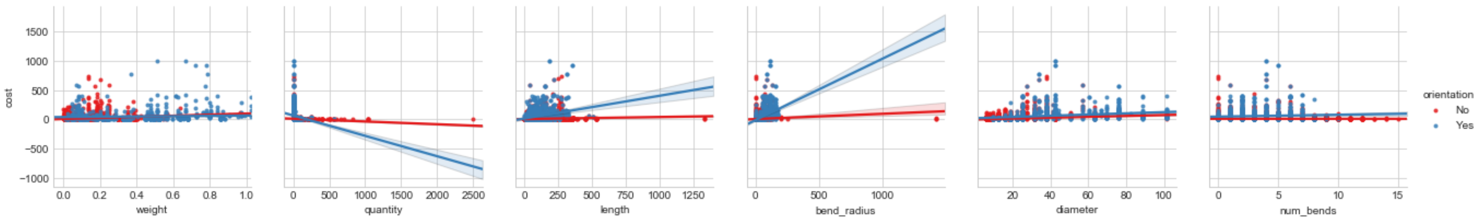
- Cost Vs Tube Specifications by Unique Features:



OBSERVATION:

1. Assemblies with Unique features had lower cost increase by weight.
2. Cost of unique features was not impacted by increase in length, bend radius or number of bends.

- Cost Vs Tube Specifications by Orientation:

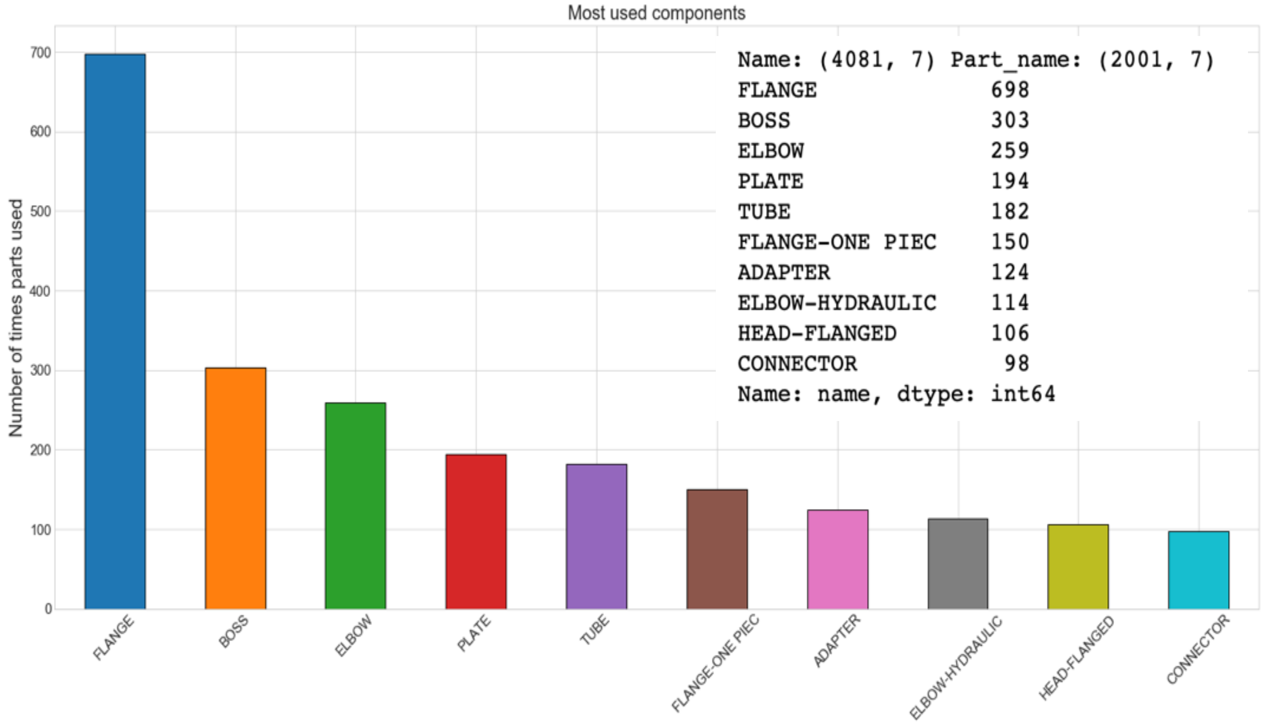


OBSERVATION

1. Assemblies with orientation had steep cost decrease when purchased in bulk.
2. Assemblies which were oriented and had Longer length and higher bend radius were costlier.
3. Cost of oriented assemblies was not impacted by assembly weight and tube diameter.

Most Used Components in Supplier Manufacturing:

This information can be useful in comparing total cost across various suppliers, and negotiating a lower unit price based on volume.
Some of the most commonly used parts found in tube assembly manufacturing were flanges, boss, elbow, plates, tube as shown in the distribution.

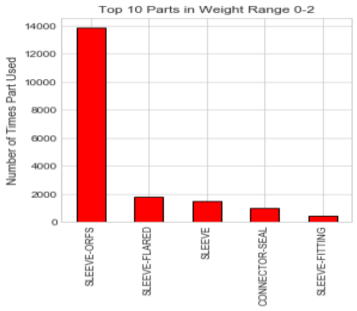
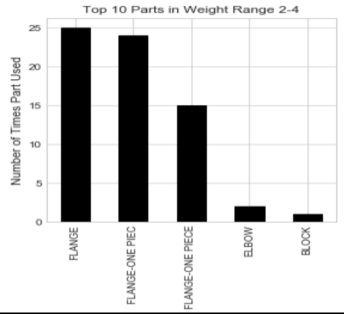
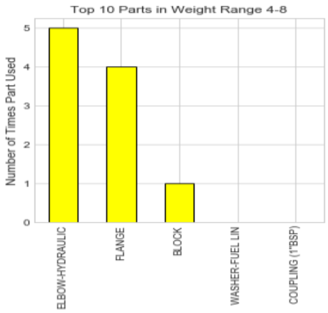
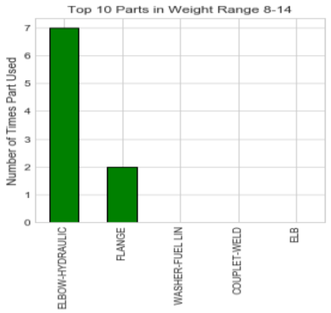
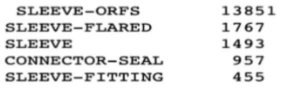
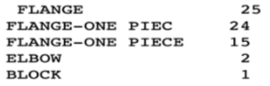
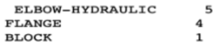
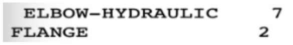


```
Name: (4081, 7) Part_name: (2001, 7)
FLANGE              698
BOSS                303
ELBOW               259
PLATE               194
TUBE                182
FLANGE-ONE PIEC     150
ADAPTER             124
ELBOW-HYDRAULIC     114
HEAD-FLANGED        106
CONNECTOR            98
Name: name, dtype: int64
```

Another important criteria that could impact cost was the weight of the assembly. As seen above in the scatter plot, heavier assemblies were costlier than lighter ones regardless of pricing type or unique features. Hence, understanding weight breakdown of most used components by the overall weight of the tube assembly would tell us which components to focus on during supplier negotiations.

As an example, assemblies which weighted under 75 percentile or 0.026/weight unit, had 3,918 unique assemblies and 15 different weights while assemblies which weighted greater than 75 percentile or 0.026/weight unit, had 1,684 unique assemblies and 242 different weight sizes. It was observed that heavier assemblies had more weight variations and fewer order transactions than lighter assemblies which had fewer weight variations and were purchased more often.

Below is the breakdown of most used parts by assembly weights:

Weight 0-2	Weight 2-4	Weight 4-8	Weight 8-14
			
SLEEVE-ORFS 13851 SLEEVE-FLARED 1767 SLEEVE 1493 CONNECTOR-SEAL 957 SLEEVE-FITTING 455	FLANGE 25 FLANGE-ONE PIEC 24 FLANGE-ONE PIECE 15 ELBOW 2 BLOCK 1	ELBOW-HYDRAULIC 5 FLANGE 4 BLOCK 1	ELBOW-HYDRAULIC 7 FLANGE 2

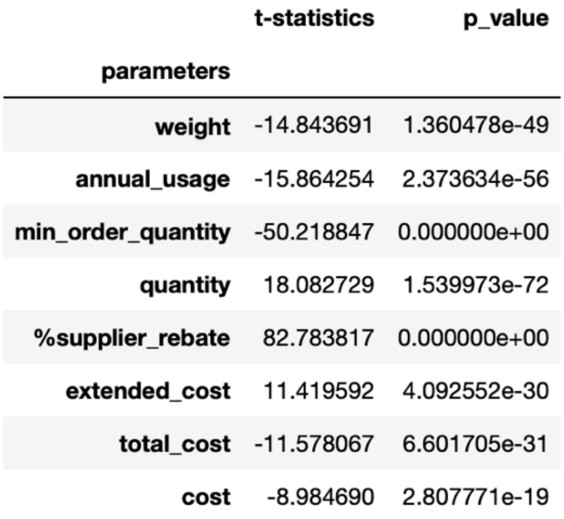
Bulk and Non-Bulk Assembly Statistical Inference:

Main objective of statistical inference was to understand the difference between bulk and non-bulk assemblies supplied in various specifications (example: weight, bend radius, area) and order patterns (example: demand, cost/unit, total cost) using statistical measures (mean and standard deviation)

The goal was to test hypothesis using two-sided t-test and confirm the significance of the result by calculating p-values of various features for both bulk and non-bulk assemblies.

If the p-value was found less than the 'significance level' of 0.05, then null hypothesis would be rejected, otherwise accepted for p-values greater than 0.05 demonstrating both assemblies features were same.

In this case, it was observed that p-values for both bulk and non-bulk assemblies were found less than the significance level of 0.05. This confirmed that both assemblies (bulk and non-bulk) differ in features and order patterns.



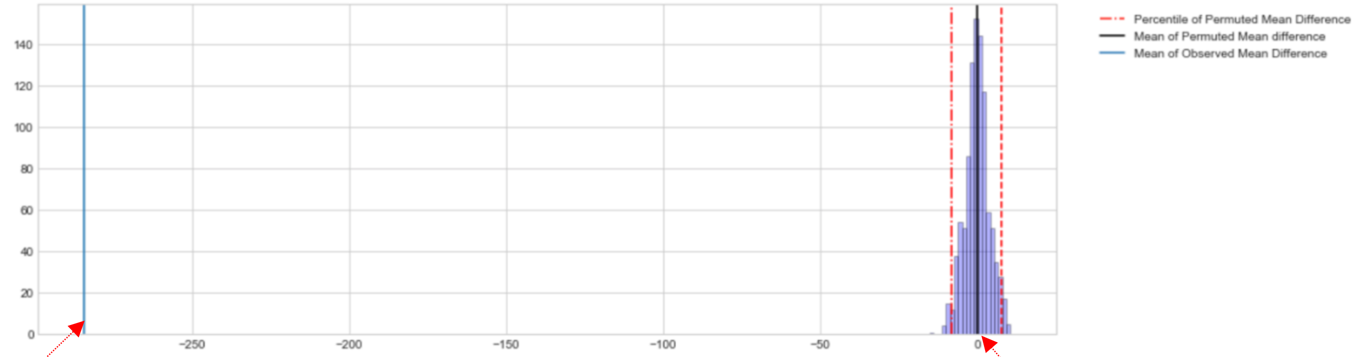
	t-statistics	p_value
parameters		
weight	-14.843691	1.360478e-49
annual_usage	-15.864254	2.373634e-56
min_order_quantity	-50.218847	0.000000e+00
quantity	18.082729	1.539973e-72
%supplier_rebate	82.783817	0.000000e+00
extended_cost	11.419592	4.092552e-30
total_cost	-11.578067	6.601705e-31
cost	-8.984690	2.807771e-19

Furthermore, by looking at the data, we know that Caterpillar purchased assemblies in both pricing types, bulk as well as non-bulk and this trend will likely continue in future and more new data will get added. Hence, could there be a possibility that business stops buying assemblies in bulk form and instead switch over to non-bulk buying i.e. just buy minimum order quantities to build equipment as on-demand than holding access bulk quantities, and if yes, how much change in p-value would there be from the existing business model.

To build this hypothesis, we collected 1000 permuted samples of bulk and non-bulk assembly features to simulate this environment and calculated mean of mean difference between all observed parameters from the given dataset and its permuted samples that we collected.

If the p-value of the mean of mean difference between permuted samples and observed assemblies was found less than 0.05 then existing trend would not continue and business would likely switch over to minimum order buying. Otherwise, null hypothesis would hold true and bulk buying would continue.

This graph shows the gap between the two mean values and its calculated p-value:



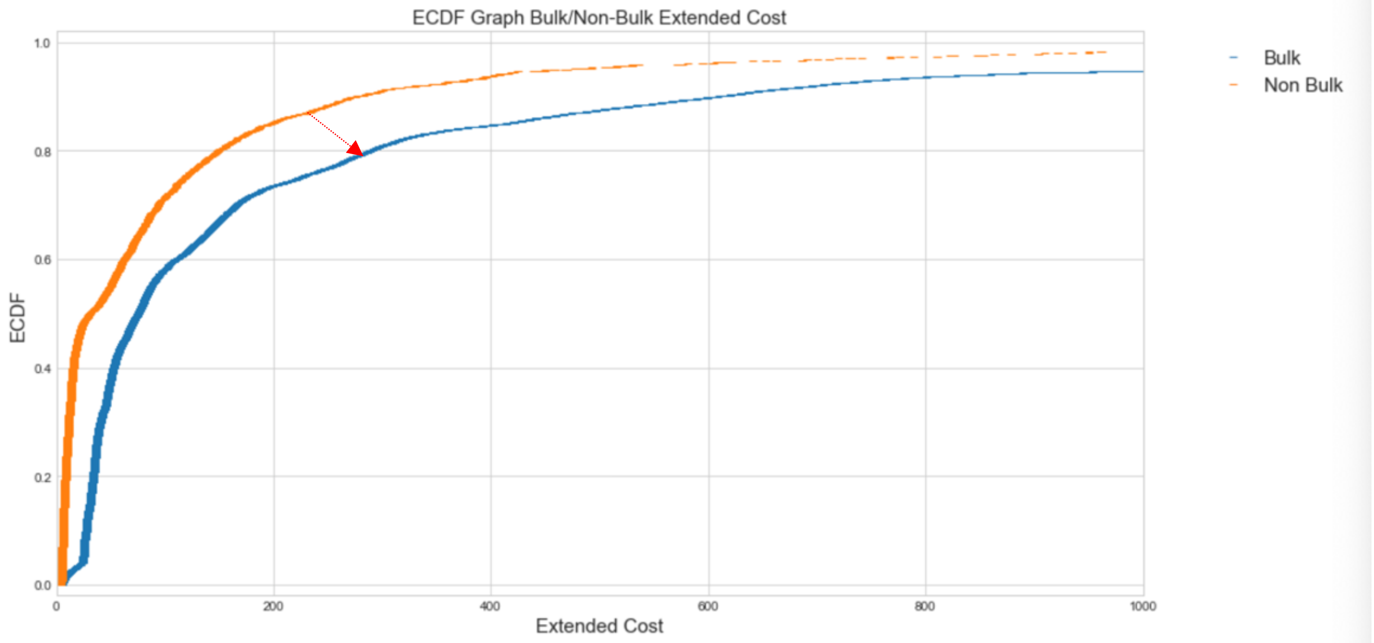
Mean of Observed Mean Difference = -284.16

p-value was equal to 1.0, greater > 0.05 hence null accepted.

Mean of Permuted Mean Difference at 95% confidence level = [-8.1, 7.93]

As per above, when the observed value of -284.16 goes beyond the upper range of permuted value of 7.93, null hypothesis would get rejected and bulk buying would discontinue. This would suggest a significant change in assembly features such as specifications and order patterns to occur, which can be monitored as more data gets collected in real time.

Let's examine how this change would be reflected in assembly cost by plotting empirical cumulative distribution function.

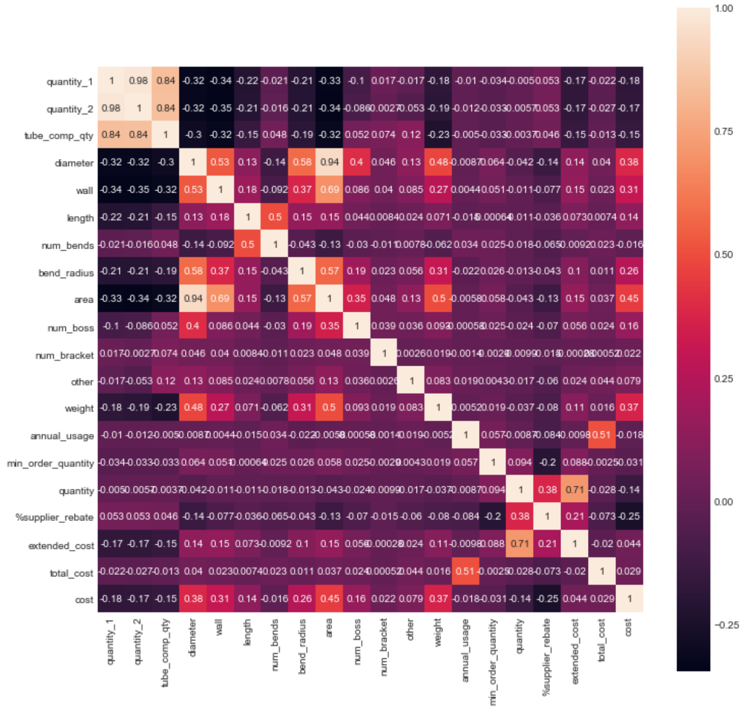


As observed in EDCF curve, bulk extended cost was found lower than the non-bulk extended cost as represented by blue line.

For switch to happen, buying non-bulk assemblies would have to cost lower than bulk i.e. the blue line would have to grow above the orange line as one of the parameters for business to review their decision. This would include comparing total cost of holding excess bulk to build equipment with just in time shipments of smaller size coming more frequently from nearby suppliers as needed by operations. (Please note, just in time data was not available to further verify this hypothesis).

Check for Multi-Collinearity:

To remove bias from the model, we plotted heat map to look for covariance values that were closer to 1 or -1, and removed features such as 'quantity_1', 'quantity_2' correlated to ‘tube_comp_qty’ with covariance value of 0.84; 'diameter' and 'wall' correlated to tube area with covariance value of 0.98 and 0.69; ‘quantity’ and ‘extended cost’ correlated with covariance value of 0.71.

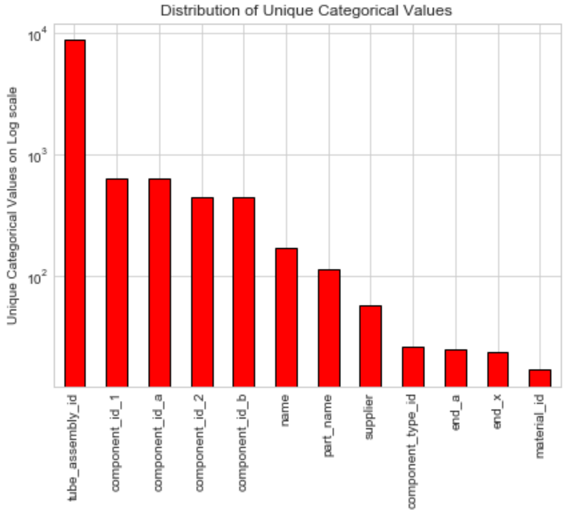
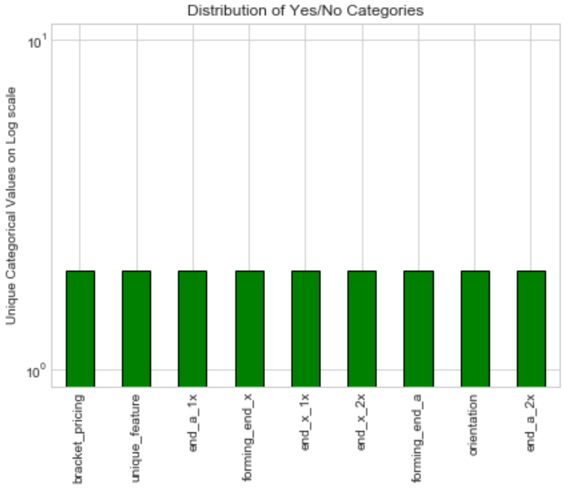


4. Results and In-depth analysis using machine learning

The goal of machine learning here is to help us predict supplier price and identify group of assemblies and suppliers who can be classified as a benchmark based on various business needs.

Predicting Supplier Price:

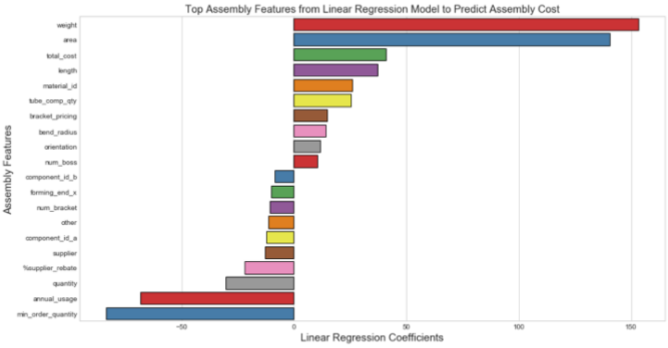
After pre-processing data and conducting exploratory data analysis, the final table contained two types of categorical values as shown in the graph below:

Ordinal represented in red	Nominal represented in Green
	

It was important to convert these values into binary values (0 ,1) so that model could also learn from categorical features. I used label-encoding (Model A) and binary encoder (Model B) to train and evaluate linear and non-parametric models.

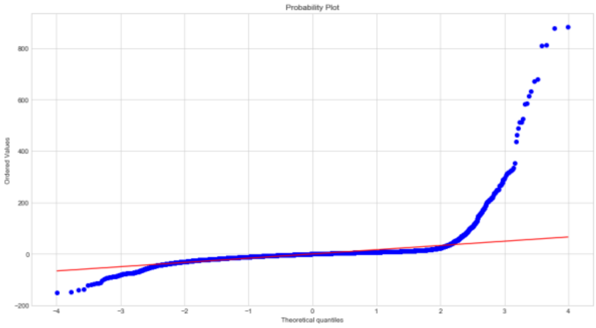
To convert categorical features, I tried one-hot encoding with label encoder but it resulted in a sparse matrix with more than 1000 + features, causing curse of dimensionality. Then tried label encoder without one-hot encoder a.k.a Model A which reduced column labels down to 36 but caused ordinal issues in the model. Finally, Binary Encoder a.k.a Model B seemed to be the best choice because it reduced 1000+ features down to 92 as well as managed ordinal issues by improving model performance.

Applied Models and Evaluation:



1. Linear Regression using Model A: This is one of the simplest supervised learning models which used a set of independent input variables (x) to predict a linearly dependent outcome (y), represented by an equation y = mx +c.
   Here m and c were the coefficients, y was the supplier price we were trying to predict, x was the input features specifications, order history, demand and more. I, applied this model to uncover the relationship between tube assembly variables and assess which features could be used to predict supplier pricing.
   As seen in the graph, we uncovered linearly positive variables such as weight, area, total cost and linearly negative variables such as supplier rebates, quantity and usage as top features in predicting supplier price with train-test evaluation score of 31% to 32%.

   However, this model also showed right-skewed residual distribution plot, indicating strong presence of non-linearity in the dataset, making it not suitable for this use case.

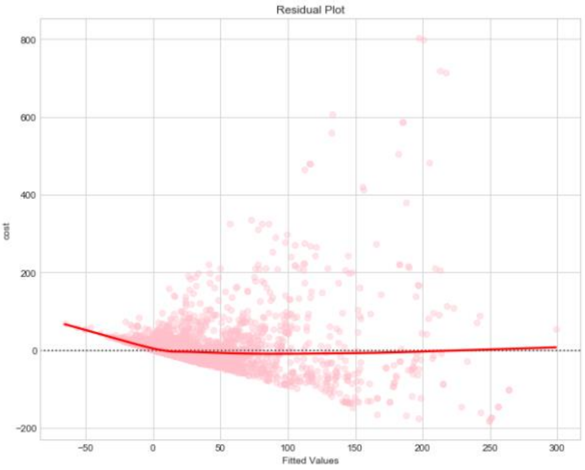


   It was observed from Quantile plot, that residual contained multiple outliers in the dataset which was either above or below the redline but not directly on the straight redline as shown in the graph.

   So, the main question arises, if we removed residual outliers, would this linear model work?

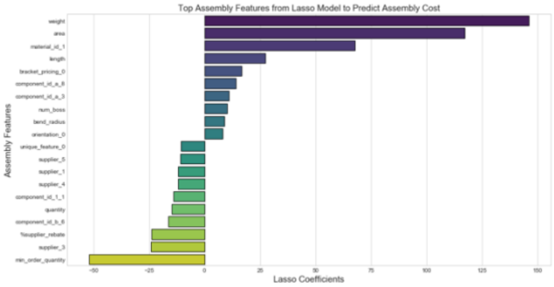
   The answer was No. After removing outliers and influence points (outliers with maximum impact on linear coefficients), R2 score (explained variance) for training and test predictions resulted in 100% accuracy, causing the model to overfit.

2. **Linear Regression using Model B:** This time, by applying binary encoder, the same model Linear Regression model which we used above, resulted in higher R2 score (40% to 42%) for training and test model predictions, accounting for 31.25% improvement over Model A. However, residual distribution continued to show right-skewness making both models not a good selection for predicting assembly cost.



3. **Lasso Regression using Model B:** This model reduced complexity of linear models by only keeping fewer features which had non-zero coefficients and were non-collinear. I applied a tuning parameter 'alpha' which penalized the sum of absolute values of the coefficients, thereby rejecting coefficients that were zero value or less important in making predictions

   I, used this model to identify important features and see how it would compare with the previous models.



   I, observed that the top three features weight, area and minimum order quantity maintained its position ranking but other features such as material id, length, contract pricing had gained higher preference in feature selection.

   R2 score improved from 32% to 41% and predictors such as weight and minimum order quantity showed 4.57% and 38% decrease in assembly cost with one-unit increase, respectively.

5. **Experiment (Linear Regression on Aggregate Table using Model A):** Out of curiosity, instead of using individual points, I wanted to build a summary table containing mean and count of numerical and categorical features. This, however, did not perform well and resulted in 11.2% difference between training and test model evaluation but on the positive note, gave lower log RMSE value.

6. **Decision Tree Regressor:** This model used both linear and non-linear features for performing classification and regression analysis. It built simple decision-based rules providing inference from the data which helped in reviewing top features that should be required to make a predictive model for this use case.

   Compared to linear models, decision tree significantly reduced log root mean squared error (RMSE) from 0.74 down to 0.19 with almost 100% accuracy and 5.3% lag between training and test evaluation sets.

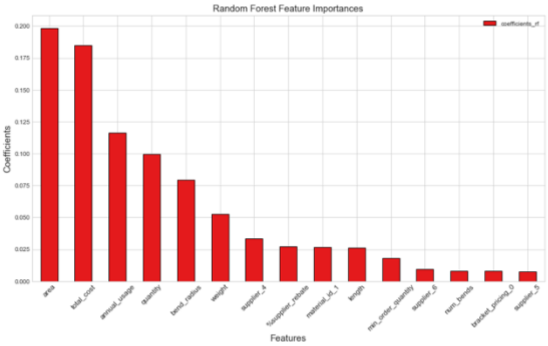
   Such high accuracy without any hyperparameter tuning could be causing decision tree to overfit the model, which is one of the problems in this model, as it continues to build tree-like structure until the last leaf is remaining (i.e. there are no decision nodes left)

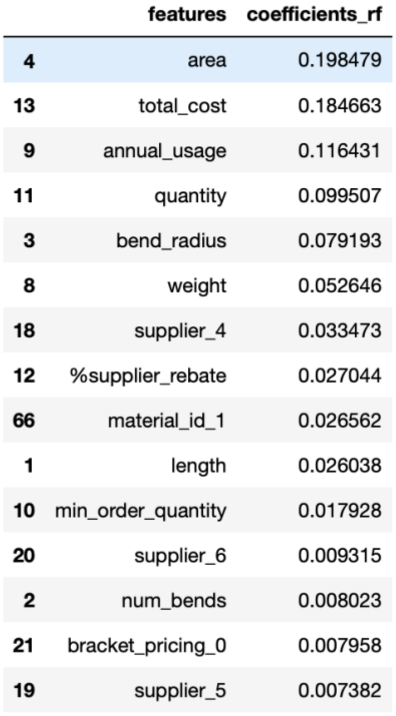
7. **Random Forest Regressor**: Unlike decision trees, Random forest consisted of a large number of individual decision trees which were majority voted or aggregated to build classification or regression models. The randomness in the model was due to random selection of subset features a.k.a bootstrapping with replacement, and random selection of decision trees of smaller sizes.

   To optimize this model, I used randomized search cross-validation and grid search cross-validation to approximate and finalize hyperparameters. My goal was to identify ideal number of decision trees with maximum depth and minimum node-leaf structure, so that model does not overfit or underfit and provides best model performance.

   During hyperparameter tuning, I observed that by not bootstrapping samples, the model used full dataset to build multiple decision trees, which took longer processing time (>30 minutes) and gave 100% accuracy, similar to decision tree model earlier.

   Hence, in trial 2 random and grid search CV, with bootstrapping sampling parameter set to True, Random Forest predicted top 15 features that made up 89% of supplier pricing prediction with 98% accuracy. The log RMSE was slightly higher from 0.19 to 0.21 and lag between train-test evaluation score was lowered from 5.3% to 5.1%.



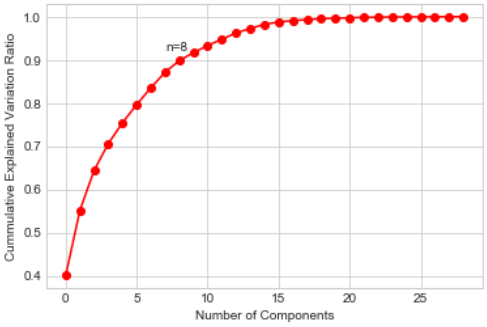


	features	coefficients_rf
4	area	0.198479
13	total_cost	0.184663
9	annual_usage	0.116431
11	quantity	0.099507
3	bend_radius	0.079193
8	weight	0.052646
18	supplier_4	0.033473
12	%supplier_rebate	0.027044
66	material_id_1	0.026562
1	length	0.026038
10	min_order_quantity	0.017928
20	supplier_6	0.009315
2	num_bends	0.008023
21	bracket_pricing_0	0.007958
19	supplier_5	0.007382

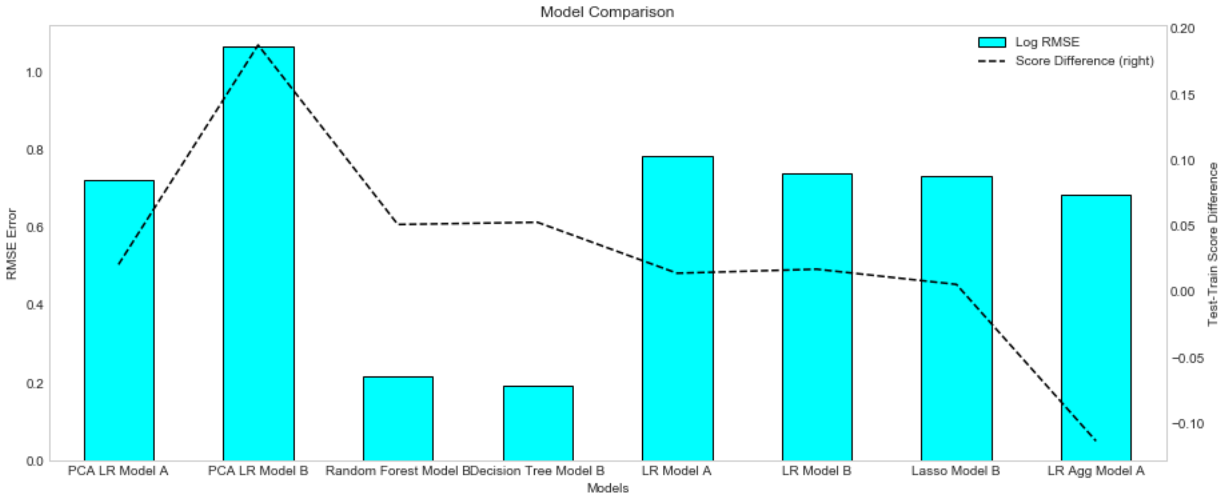
These top features gained much higher preference than what we obtained from linear regression models earlier, further indicating that linear model coefficients were indeed impacted by residual outliers and influence points.

Total variance explained by top 15 features: 0.894642

8. **PCA with Linear Regression using Model A & Model B:** Finally, I tried dimension reduction technique (PCA) by identifying optimal number of principal components that explained maximum variance in the supplier price and passing it through linear regression model, but it resulted in lower training and test evaluation score and high Log RMSE values, similar to aggregation model experimented earlier.



Predictive Model Summary and observations:
This is the summary of all linear models tried so far in building predictive supplier pricing.

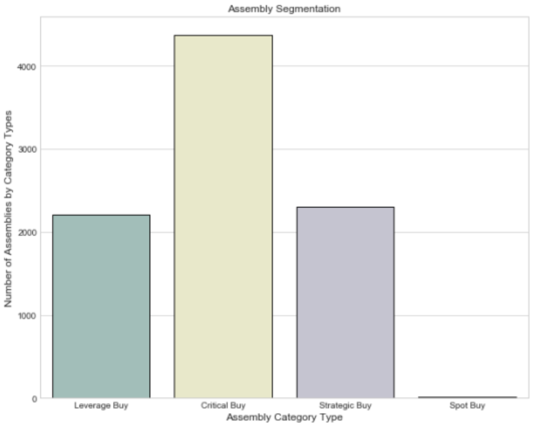
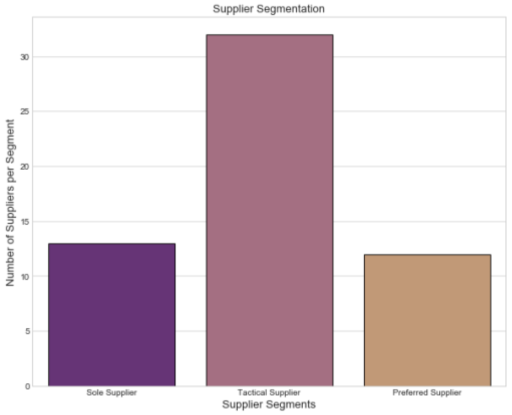


1. **Linear models:** Log RMSE values, as shown by bar, were found comparable with the exception of PCA LR on model B. This is probably because PCA prefers numerical values to find principal components and does not work well with binary encoded data, contained in Model B.

2. **Random Forest:** Overall resulted in good model accuracy 98% after hyperparameter tuning and gave top 15 features which explained 89% variance in predicting supplier price. In future, we can use these features to build linear model again using lasso or simple regression model and compare results.

Categorizing Tube Assemblies and Suppliers using Clustering:

1. **Pre-Classification**: My goal here was to establish a criteria for segmenting tube assemblies and suppliers. I, anticipated business such as finding cost efficiencies in the form of supplier rebates, building long term contractual relationships and reducing market difficulty so that assemblies could be procured from a low cost technically compliant supplier.
Below is the pre-segmented distribution of assemblies and suppliers. This would be useful in describing portfolios by business needs and justifying various clusters when we apply unsupervised learning i.e. Without target labels, to our model.

<table>
<tr><th>Pre-Assembly Segmentation: 4 Categories</th><th>Pre-Supplier Segmentation: 3 Categories</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>

- Assembly Portfolio Breakdown:
  - Critical Buy 49.1%
  - Strategic Buy 25.9%
  - Leverage Buy 24.8%
  - Spot Buy Negligible

</td><td>

- Supplier Portfolio Breakdown:
  - Tacticle Suppliers 56.1%
  - Sole Suppliers 22.8%
  - Preferred Suppliers 21.1%

</td></tr>
</table>

2. **Clustering**: After pre-classification was established, I applied log transformation to remove skewness in the data as well as applied standard scaler to ensure model data was on the same unit of scale. This step was crucial in ensuring that meaningful clusters were formed when model was built. After normalizing the data, I built various clustering algorithms such as K-means, Affinity Propagation and few more to group assemblies and suppliers by the business needs.
   The objective here was to benchmark these clusters and identify assemblies and suppliers who fully meet pre-set criteria. Another important aspect in analyzing these clusters was evaluating cluster quality. I, used Silhouette score as a measure, which ranges between +1 (good cluster) to -1 (poor cluster), as shown in the summary table below.

	Assembly Cluster Quality	Supplier Cluster Quality
Model	Silhouette Score	Silhouette Score
• K-Means	0.46	0.35
• Affinity Propagation	0.41	0.35
• AgglomerativeClustering	0.44	0.34
• SpectralClustering	0.07	0.36
• DBSCAN	0.33	0.14

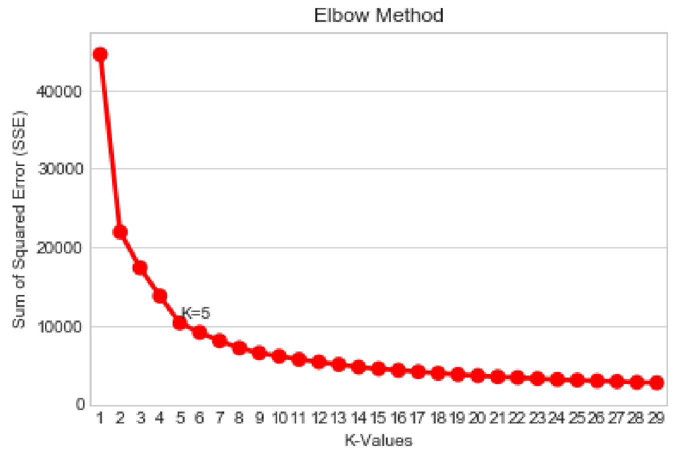
From above, K-means was selected for building assembly cluster and optimal value of K was obtained using elbow method, which gave us a point where sum of squared error values showed marginal decrease with the increase in number of clusters.

In case of supplier clusters, Spectralbiclustering and K-means showed comparable silhouette score and gave less dense quality clusters. Hence, I chose K-means as a common algorithm to compare cluster formation.

Comparison between Assembly and Supplier Clusters:

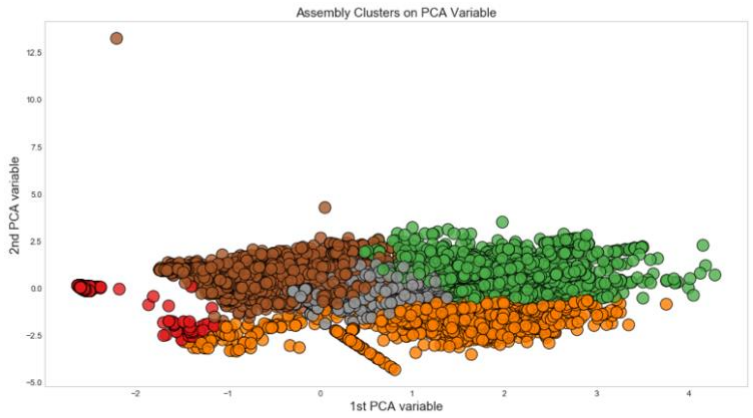
Assembly Clusters

- k-value: 5 clusters

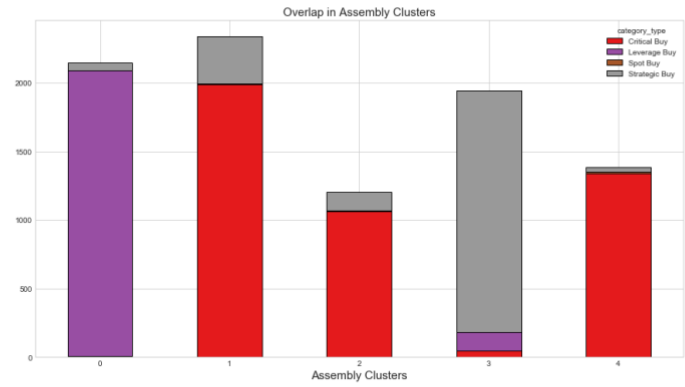


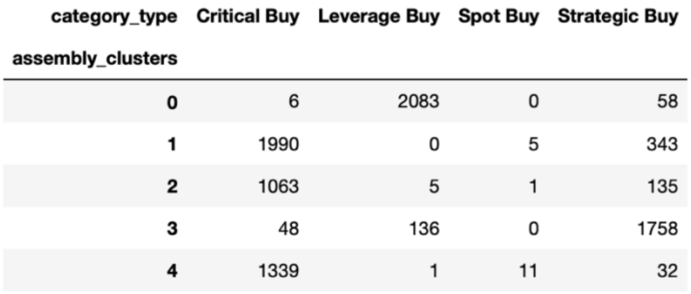
- Quality: Silhouette Score 0.46

Dense clusters on PCA variables



- Overlap: All assembly categories had some strategic buy

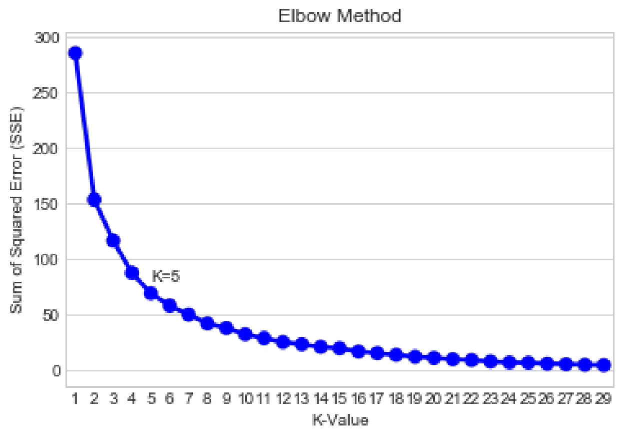




category_type	Critical Buy	Leverage Buy	Spot Buy	Strategic Buy
assembly_clusters				
0	6	2083	0	58
1	1990	0	5	343
2	1063	5	1	135
3	48	136	0	1758
4	1339	1	11	32

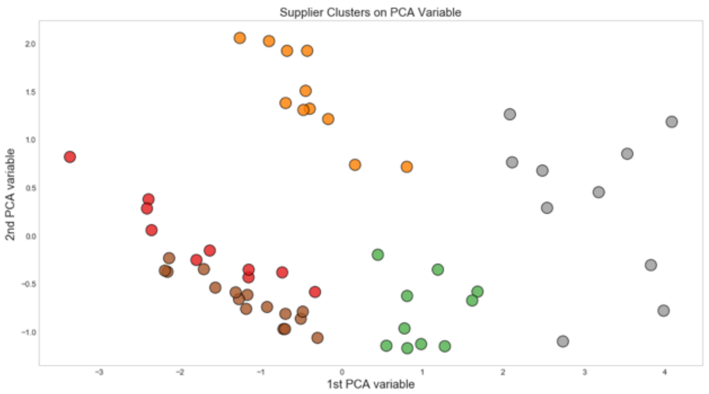
Supplier Clusters

- k-value: 5 clusters

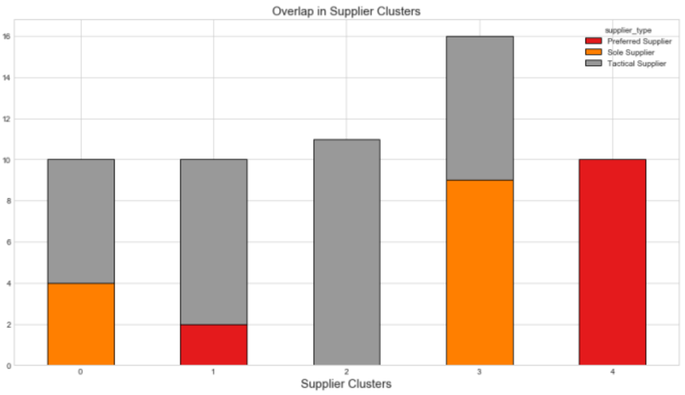


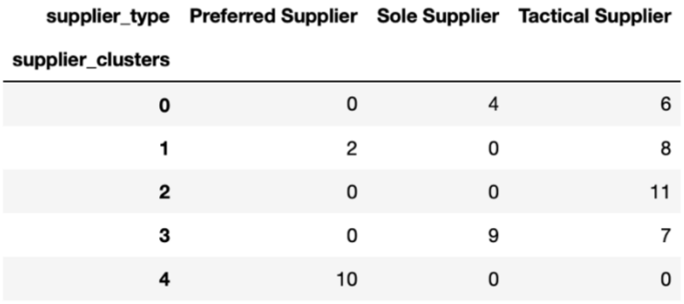
- Quality: Silhouette Score 0.35

Less Dense Clusters on PCA variables



- Overlap: Tactical suppliers had some sole & preferred suppliers





supplier_type	Preferred Supplier	Sole Supplier	Tactical Supplier
supplier_clusters			
0	0	4	6
1	2	0	8
2	0	0	11
3	0	9	7
4	10	0	0

<table>
<tr>
<td>

- Benchmarked Cluster:

Cluster 3: High rebates and fewer suppliers managing high spend

- Other Clusters:

Cluster 1,2 and 4: High supplier spend, high frequency purchase, lowest supplier rebates.
Cluster 0: No spend, offered high rebates (potential for leverage buy).

</td>
<td>

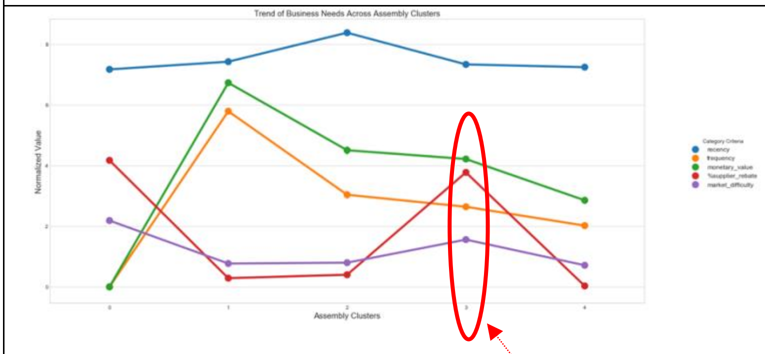
- Benchmarked Cluster:

Cluster 4: High rebates offered by highly preferred suppliers

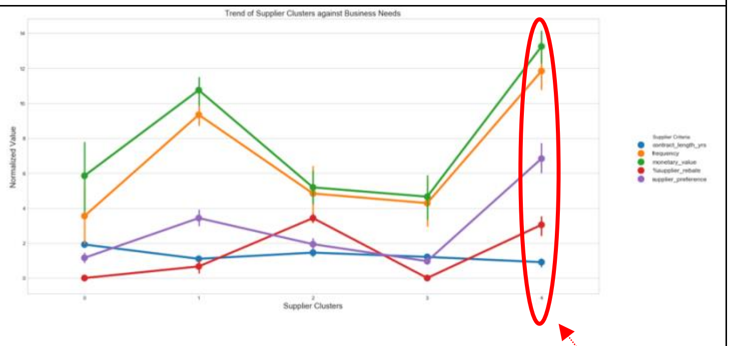
- Other Clusters:

Clusters 0, 3: Lowest supplier preference, no rebates (similar to Sole Suppliers). Opportunity for spend consolidation.
Cluster 2: Similar to cluster 4 with very low supplier preference. Opportunity to develop cluster 2.
Cluster 1: High spend new preferred suppliers with relatively low rebate. Opportunity for contract negotiation.

</td>
</tr>
<tr>
<td>



</td>
<td>



</td>
</tr>
</table>

Conclusion:

1 By using Random Forest and K-Means, we were able to establish pricing prediction and benchmark assemblies and supplier clusters from a pool of 8,855 unique assemblies and 57 suppliers.
2 Out of multiple assembly features and observations, we were able to select out top 15 assembly parameters (both linear and non-linear) that explained 89% variance in the predicted supplier price with 98% accuracy.
3 By choosing optimal k-value, cross tab count and silhouette score, we were able to evaluate quality of grouped clusters and identify key business trends and opportunities.

Future Possibilities:

1. Monitor Supplier Performance: We can add other features such as on-time delivery, safety statistics, contract compliance and inventory levels to gain further insights about supplier performance.
2. Build Real Time Dashboards: We can connect our model to real time data feed and collect actionable insights on the fly.
3. Further Modeling: We can try other algorithms such as time-series to predict annual supplier pricing.

----End of report----